Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 13

Birth-Death Queues: General Theory, M/M/1 Queues and their Steady State Solution

Hi and hello, everyone. After having seen the basics of queues and some general results and having acquired the background from the theory of stochastic processes needed to understand and analyze a queueing system through a Markov chain model. We will now move into the analysis of queueing systems of various different models which can model different queueing systems. The first step in the analysis is to consider queueing systems, broadly called Birth-Death queueing systems or elementary queueing systems, or exponential model-based, exponential distribution-based queueing systems, the simplest ones.
Many simple but interesting and nontrivial queueing systems can be studied through birth-death processes. It is so simple yet highly effective in implementing many of the real-life situations that can be modelled through these birth-death processes. Recall

- A BDP is a continuous-time Markov chain on the state space of non-negative integers, where the state transitions occur only to neighbouring states.
- From a queueing perspective, an arrival is regarded as birth and a departure is regarded as death with the state denoting the number in the system.

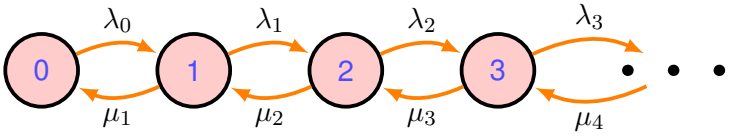
  Whenever we are talking about a solution of analysis of a queueing system, or we primarily mean the number in the system so, that will be our main objective or main object of study in any queueing system that is queueing system whether it is single or network or anything that is with the primary objective will be there. So, that is why the state mainly denotes the number in the system.
- In a birth-death queue, the arrivals as well as the departures happen one-at-a-time.

  So, the probability of arrival and departure happening exactly at the same time is 0, or negligible. So, we can always slot it in such a way that they are distinct.
- We will always assume (unless stated otherwise) that the arrivals occur from an infinite source, for all queueing systems (birth-death or otherwise).
  when we will go later on into more general Markovian queueing systems or semi Markovian queueing systems, we will always assume that the arrivals occur from an infinite source which means that there is an infinite population from which the arrivals happen to the particular queueing system which is what we want to study. Wherever that is a finite source, we will mention that as well. So otherwise, we will assume that this is the case.

- We will mainly be concerned with the steady-state (or equilibrium) analysis of the queueing systems (though we will see briefly some transient analysis as well).
- The birth-death queueing models are also sometimes referred to as 'exponential models'.

Again, we will repeatedly come back to the birth-death process because this is the last time we will be doing it after that; we are not going to come back again.

- A birth-death process (BDP) is a continuous time Markov-chain (CTMC) on state space $\{0, 1, 2, 3, \dots\}$, with state transitions occurring as unit jumps up or down from the current state.
  - When $n \geq 0$, time until next arrival (or "birth") $\sim Exp(\lambda_n)$.
    ▶ At arrival, the system state moves from n to $n+1$ at rate λ_n.
  - When $n \geq 1$, time until next departure(or "death") $\sim Exp(\mu_n)$.
    ▶ At departure, the system state moves from n to $n-1$ at rate μ_n.
  - Here the 'states' denote the number of customers in the system.



  - The generator matrix (or rate matrix) for a BDP is $Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$.

- Let p_n denote the long-term fraction of time the system is in state n.
- p_n can be determined from $\mathbf{0} = \boldsymbol{p}Q$, subject to certain conditions on λ_n and μ_n.
- For the BDP, the vector-matrix equation $\mathbf{0} = \boldsymbol{p}Q$ can be written in component form as

$$\begin{aligned} 0 &= -(\lambda_n + \mu_n)p_n + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \quad n \geq 1 \\ 0 &= -\lambda_0 p_0 + \mu_1 p_1 \end{aligned}$$

  Equivalently,

$$\begin{aligned} (\lambda_n + \mu_n)p_n &= \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \quad n \geq 1 \\ \lambda_0 p_0 &= \mu_1 p_1. \end{aligned} \tag{1}$$

- The above equations in (5) are called "*global*" balance equations which must be satisfied at equilibrium, since they equate the total mean flow into each state with the total mean flow out of that state.

This is called the "global" balance equation or simply balance equation; we may not use the word global I mean often, but balance equation we always mean the global balance equation, which is what is given here. Now, how do you write down this again, you can look at this graph.

- One method (an iterative method) of finding the solution for p_n is described below.
  Rewriting (5), we get

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \quad \text{and} \quad p_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} p_n - \frac{\lambda_{n-1}}{\mu_{n+1}} p_{n-1}, \quad n \geq 1.$$

  From the above,

$$p_2 = \frac{\lambda_1 + \mu_1}{\mu_2} p_1 - \frac{\lambda_0}{\mu_2} p_0 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_0} p_0 \qquad (\text{by putting the value of } p_1)$$

  Similarly,

$$p_3 = \frac{\lambda_2 + \mu_2}{\mu_3} p_2 - \frac{\lambda_1}{\mu_3} p_1 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0.$$

  In general, one can show by mathematical induction that

$$\boxed{\begin{aligned} p_n &= \frac{\lambda_{n-1} \lambda_{n-2} \ldots \lambda_0}{\mu_n \mu_{n-1} \ldots \mu_1} p_0, \qquad n \geq 1 \\ &= p_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \end{aligned}} \tag{2}$$

- Since probabilities must sum to 1, i.e., $\sum_{n=0}^{\infty} p_n = 1$, it follows that

$$p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1} \tag{3}$$

  We observe that $p_0 > 0$ if and only if $1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}$ is convergent and this is the necessary and sufficient condition for the existence of a steady state distribution $\{p_n\}$ for the BDP because this is the condition that makes the underlying BDP positive recurrent, and in such case, it is already irreducible, positive recurrent, and hence there is a unique stationary distribution which is also same as the limiting distribution which is given by the stationary equation $pQ = 0$ and $\sum_{n=0}^{\infty} p_n = 1$.

- The "product" solution obtained above for $p_n, n = 0, 1, 2, \ldots$, is a *principal* equation in birth-death queues, and is extremely useful in analyzing a variety of queueing models.

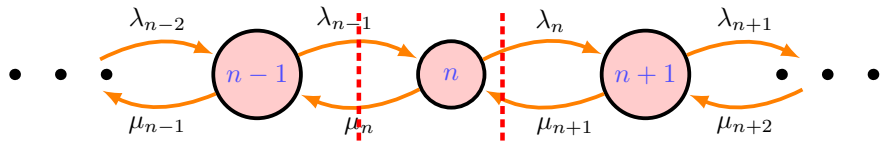
Once you know the rate because here we remember, we are given in terms of generically λ_n's and μ_n's. So, now, once you know the rates, then you can obtain the solution very nicely by simply putting it here by simply p_n's in terms of $p_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1} p_0, \quad n \geq 1$. So, this is the main equation for the birth-death in the iterative method; this is what the iterative method; this is the main equation, and once p_n's expressed in terms of p_0, then $\prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}$ is also getting get simplified. So, $p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1}$ series and under the condition whatever be the condition under which this series is convergent that is the condition for the positive recurrence of the chain, and hence one can directly use it. So, $p_0 = \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right)^{-1}$, so, one can directly use. So,

that is why this is a kind of a principal equation for birth-death queues because once you have birth-death queues, you can simply substitute that is going to be one of the ways to solve; now you do not need to write down

$$0 = -(\lambda_n + \mu_n)p_n + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \quad n \geq 1$$
$$0 = -\lambda_0 p_0 + \mu_1 p_1$$

flow balance equation and start from this to get solved. Once you know that it is a birth-death queueing system, simply determine the rates λ_n's and μ_n's. Substitute in $p_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1}p_0, \quad n \geq 1$ to get p_n in terms of p_0, and the exact same thing will give you what would be this $p_0 = \left(1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}\right)^{-1}$. Once we get p_0, then all p_n's are determined, and you are done with the solution that is how you know things will go, and that is why this will be the principal equation for birth-death queues in analyzing the variety of queueing models which can be modelled through a birth-death process.

- A second method of obtaining the above product solution is through the *local* balance equations $\lambda_{n-1}p_{n-1} = \mu_n p_n, \ n \geq 1$.



Now, look at this

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \quad n \geq 1 \tag{4}$$
$$\lambda_0 p_0 = \mu_1 p_1.$$

Now, take the first equation with $n = 1$; now, what will happen then. You have $(\lambda_1 + \mu_1)p_1 = \lambda_0 p_0 + \mu_2 p_2$. On the left side also, there is $\mu_1 p_1$, and this side is, let us say, $\lambda_0 p_0$, and from $\lambda_0 p_0 = \mu_1 p_1$, we know that these two quantities are equal. So, this gets cancelled, and you are left with $\lambda_1 p_1$, which is equal to $\mu_2 p_2$. $lambda_0 p_0 = \mu_1 p_1$ and $\lambda_1 p_1 = \mu_2 p_2$. Now, one more step you go, you will get $\lambda_2 p_2 = \mu_3 p_3$ and so on. So, basically, from

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \quad n \geq 1 \tag{5}$$
$$\lambda_0 p_0 = \mu_1 p_1.$$

you will get that equation which is what we call as "local" balance equation, and that is what this $\lambda_{n-1}p_{n-1} = \mu_n p_n, \ n \geq 1$ is true; this is what you are seeing here. Now, from $\lambda_{n-1}p_{n-1} = \mu_n p_n, \ n \geq 1$, you can easily see $p_n = \frac{\lambda_{n-1}p_{n-1}}{\mu_n}$, and you iterate it, what you will get is $p_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1}p_0, \quad n \geq 1$. That is what you are; you are going to get it, and this is what we call the local balance equation.

Now, what is this local balance equation? Remember, this global balance equation we wrote with respect to state right looking at the state and when what is the flow into this particular state and out of the particular state will give you the global balance equation or simply balance equation. So, one can write down local balance equations directly for birth-death queueing models, but this is not true for any general Markov chain. The global balance will always imply local balance, but this local balance need not always imply the global balance; and there is another concept called

reversibility, and if the chain is reversible, of course, this both imply each other, but in this case, we do not need to worry because we will always handle with the global balance equations or simply balance equations or flow balance equations. But for birth-death, there is a simpler way of obtaining the flow balance, which is via this local balance, and that is you are looking at as if you are standing somewhere in the middle, not on the state, but somewhere between two states and you are looking at the flow on the one side and the flow in the other side and in equilibrium, they must be equal that gives you the local balance equation. This is only a piece of additional information we are giving, but we will not use this local balance in any of our analyses. We will always write global balance equations. Now, consider having understood BDP again once more with a little bit more details.

So, we will now consider the first basic model, which is called $M/M/1$ queues.

- The celebrated $M/M/1$ queue is the simplest nontrivial interesting queueing system because of randomness you have brought in.
  If it is deterministic, then, of course, that suppose in our notation what we mean this M denote the interarrival times the second M denotes the service time and 1 server with the infinite capacity and FCFS discipline. Now, suppose if these are all deterministic, then there is nothing much to analyze; everything will fall through, but the moment you bring in randomness, the simplest model is basically $M/M/1$ model. So, $M/M/1$ is in Kendall's notation; we just explained what that means.

- The study of behaviour of an $M/M/1$ system is vital on many counts.
  ▲ Serves as a benchmark for new methods and new performance measures.

  So, first of all, anything you develop and you try to apply for an $M/M/1$ queueing system to see that things work well or what improvements it gives or whatever is the case that you are looking for it. So, this, in some sense, is a benchmark model because anything that you want to do, you always try to do at the level of $M/M/1$ queue before you move to any other queue, say $M/G/1$ or anything of that sort.

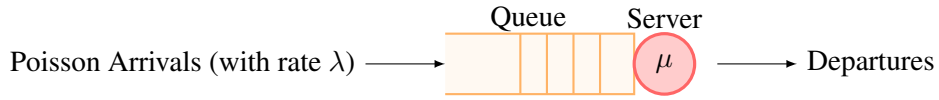
  ▲ Much of the behaviour manifested in $M/M/1$ is characteristic of more complex queueing system behaviour.

  So, the behaviour that you observe here is what would be similar to what you would get to observe in more complex systems. So, the major aspects can of any particular complex queueing system be heard if you analyze the M M 1 version of it M M 1 type. So, that is why the understanding of this M M 1 queue is important it is very simple it is not at all complex. But you have to look at it in a way whether you know what is that it is giving you what is that that you want to take out of the analysis of this. So, you always have to look for it to get what is the nature of this behaviour and so on. So, that is why the study of this simplest model is what is very important. Now, what is it M M 1 queue?

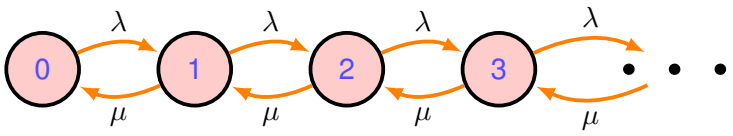
- An $M/M/1$ queue is a single server queue with a Poisson arrival process and with an exponentially distributed service time distribution.
  ♦ Interarrival times are assumed to be IID $Exp(\lambda)$ distributed, service times are assumed to be IID $Exp(\mu)$ distributed and they are independent of each other.
  ♦ The arrival rate $\lambda > 0$ and the service rate $\mu > 0$ are fixed.

- It is further assumed that an infinite waiting space is provided and that the customers are served according to a FCFS (or FIFO) discipline.

So, in pictorially like, if you want to depict this particular queueing system, this is what you will have.



- Let $N(t)$ denote the number of customers in the system at time t.
- The process $\{N(t), t \geq 0\}$ can be modelled as a BDP with $\lambda_n = \lambda, n \geq 0$ and $\mu_n = \mu, n \geq 1$.
  The transition rate diagram for the BDP (or for the $M/M/1$ queue):



Now, you can also look at what is the Q matrix here in this particular case.

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ \mu & -\lambda-\mu & \lambda & 0 & \dots \\ 0 & \mu & -\lambda-\mu & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Now, the quantity of interest $p_n(t)$ here whenever we have such a system. So, now, you assume that things are interarrival. So, this basically the conditions that we have put for $M/M/1$ queue that you know in a small interval of length, say delta t what the probability of arrival is, what is the probability of a departure, what is the probability of more than one event happens that is how you typically define for a BDP. And exactly the same thing is happening here for the $M/M/1$ queue because of your assumption that the interarrival times are exponential, service times are exponential. So, basically, you are getting exactly the same condition, and hence this can be modelled through a birth-death process because it is the transition that happens one at a time. So, the quantity of interest here you will be for this queueing system is the number in the system at time t.

- Let $p_n(t) = P\{N(t) = n\}$ for $n \geq 0$. The system of differential-difference equations (forward Kolmogorov equations) for the system state probabilities this BDP are given by

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t) \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu) p_n(t) + \mu p_{n+1}(t), \quad n \geq 1. \end{aligned}$$

- Ideally, we would seek a solution (known as time-dependent or transient solution) to the above equations, subject to the initial probability distribution $\{p_n(0), n \geq 0\}$.

- For now, we will focus on the equilibrium behaviour of this system model which means that we assume that the system has reached an equilibrium because the system is in a stable situation and has reached a steady-state situation. So, it is operating under a steady-state situation, and in that situation, we want to look at what are these quantities; that means, we try to determine $p_n = \lim_{t\to\infty} p_n(t)$ directly rather than obtaining $p_n(t)$ and then taking its limit.

We said that complexity is there, but we will see how complex is that as an example for $M/M/1$ queue a little later, but in all our analyses except that part when we are going to the transient analysis, the rest of the complete course we are dealing with only equilibrium analysis. So, we will always be interested in what happens to p_n and how we can get p_n which is what is of our interest.

- Assume that the $M/M/1$ system is in steady state.
- The (global) flow balance equations for this system are

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda+\mu)p_n &= \mu p_{n+1} + \lambda p_{n-1}, \qquad n \geq 1. \end{aligned} \tag{6}$$

  Equivalently,

$$p_1 = \frac{\lambda}{\mu} p_0 \quad \text{and} \quad p_{n+1} = \frac{\lambda+\mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1}, n \geq 1. \tag{7}$$

- Three methods to solve the above system of equations are:
  ▶ Iterative method, Generating function method and Operator method.
  ▶ The purpose is to illustrate the methods, as one method may be superior/easy to other depending on the model at hand.

So, you will try to adapt which method is going to work; that is why I said these are the not just all the methods; there are many more different methods that can be utilized, but we will highlight these three methods for the time being because these are the things which we will be using it in the future. The reason is just for illustration purposes of how these methods work.

- Since $M/M/1$ system is a birth-death process with $\lambda_n = \lambda$ and $\mu_n = \mu$, we can directly apply the iterative method given earlier for a BDP with $\lambda_n = \lambda$ and $\mu_n = \mu$.
- It follows that

$$p_n = p_0 \prod_{i=1}^{n} \left(\frac{\lambda}{\mu}\right) = p_0 \left(\frac{\lambda}{\mu}\right)^n, \quad n \geq 1.$$

  We can get p_0 using the fact that $\sum_{n=0}^{\infty} p_n = 1$. This gives us

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} p_0 \left(\frac{\lambda}{\mu}\right)^n = p_0 \sum_{n=0}^{\infty} \rho^n, \text{ where } \rho = \frac{\lambda}{\mu} = r \text{ is the offered load or traffic intensity.}$$

$$\Rightarrow\ p_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1-\rho. \qquad \left(\text{using geometric sum: } \sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho}, \quad \text{for } \rho < 1\right)$$

Therefore, we obtain $p_n = P\{N = n\} = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots, \quad \rho = \frac{\lambda}{\mu} < 1.$
Thus, $N \sim Geo(1-\rho)$.

We are simply whatever iterative method that we have already developed there; we are simply employing it with $\lambda_n = \lambda$ and $\mu_n = \mu$ parameters, and we are getting the solution p_n; this is one method. The other method is using generating functions. How do we do?

- Define the probability generating function $P(z) = \sum_{n=0}^{\infty} p_n z^n$, ($z$ complex with $|z| \leq 1$).
- We solve for and express $P(z)$ as a power series and then obtain p_n's as coefficients of z^n.
- Rewrite the balance equations in terms of ρ as

$$\begin{aligned} p_1 &= \rho p_0 \\ p_{n+1} &= (\rho+1)p_n - \rho p_{n-1}, \quad n \geq 1. \end{aligned}$$

Multiply both sides of the second equation above by z^n & sum over n from 1 to ∞ to get

$$\begin{aligned} & \sum_{n=1}^{\infty} p_{n+1} z^n = (\rho+1)\sum_{n=1}^{\infty} p_n z^n - \rho \sum_{n=1}^{\infty} p_{n-1} z^n \\ \Rightarrow \quad & \frac{1}{z}[P(z) - p_1 z - p_0] = (\rho+1)[P(z) - p_0] - \rho z P(z) \\ \Rightarrow \quad & \frac{1}{z}[P(z) - (\rho z + 1)p_0] = (\rho+1)[P(z) - p_0] - \rho z P(z), \qquad (\text{using } p_1 = \rho p_0) \\ & \Rightarrow \quad P(z) = \frac{p_0}{1-\rho z} \end{aligned}$$

- We need to find p_0. Observe that $P(1) = \sum_{n=0}^{\infty} p_n 1^n = \sum_{n=0}^{\infty} p_n = 1$. Therefore, we have

$$P(1) = 1 = \frac{p_0}{1-\rho}, \quad \Longrightarrow \; p_0 = 1 - \rho.$$

Since $\{p_n\}$ is a probability distribution, $P(z) > 0$ when z is real and $z > 0$. Therefore $P(1) > 0$ and this implies that $P(1) = \frac{p_0}{1-\rho} > 0$. This means that $\rho < 1$. In summary,

$$\begin{aligned} P(z) &= \frac{1-\rho}{1-\rho z}, \quad \rho < 1, \quad |z| \leq 1 \\ &= (1-\rho)\sum_{n=0}^{\infty} \rho^n z^n. \quad (\text{using the geometric series sum}) \end{aligned}$$

Since p_n is the coefficient of z^n, we have

$$p_n = (1-\rho)\rho^n, \quad n \geq 0, \quad \rho = \frac{\lambda}{\mu} < 1.$$

So, this is the PGF approach or the generating function approach. Some observations based upon this algebraic form of this for this model can be given, which is maybe relevant. I mean, it is very simple here, but you know the similar idea is what is going to be prevalent elsewhere when you have a more complex situation.

- Some observations based on the algebraic form of $P(z)$ as given above for this model:

$$P(z) = \frac{1-\rho}{1-\rho z}, \quad \rho < 1, \quad |z| \leq 1.$$

► This expression is a rational function, being quotient of two polynomials.
The numerator is a constant and the denominator is a linear form $1 - \rho z$.
The denominator has a single zero $1/\rho$, which is the reciprocal of the traffic intensity and its value is greater than 1.

So, you will be looking for a root of that nature if we have multiple roots and if the polynomial here is not linear, but quadratic of order anything, then you will look for rows which are of having this property, this zero of denominators which is value is greater than 1 is what then will give you the required quantities. So, this is a rational function; the rational function is what is easier, whether it is Laplace transform or z transform; the rational functions are easier to deal with it because then you can make partial fraction expansions for these rational functions, and you can write down nicely. You can try to get the corresponding series expansion so that you can get the quantities there that are a nicer form that you would expect to have.

► For some models, it may easy to find an expression for $P(z)$ but it may be difficult to find its series expansion to obtain p_n's.
But $P(z)$ still provides useful information. For example, we can obtain

$$L = E(N) = \sum_{n=0}^{\infty} n p_n = \left.\frac{\mathrm{d}P(z)}{\mathrm{d}z}\right|_{z=1} = P^{'}(1).$$

Later on, you will see when we deal with a little more complex models like, at whatever stages we will be using this kind of idea. So, the L can be obtained from here even though you do not have the explicitly what is this p_n. So, from P, I can differentiate with respect to z, and I can obtain its mean of mean number in the system. So, this is some observation that you are making with respect to PGFs that ideally, like the algebraic form rational functions is what we want, and that will give us the ideas of how one can handle these PGFs and how we will get these zeros for the denominator for which will serve our purpose and even if you are not able to obtain explicit expression or not able to get from the PGF the p_n's you can still obtain some useful information. So, this is the second method that we are using.

The third method is using operators; this is much similar to the differential equation that you might be aware of because that is also an operator; this differential operator, this is a generic operator, which is what the difference operator; basically, you can say here.

- We use the theory of linear difference equations to solve for $\{p_n\}$, starting from the main equation in our flow balance equations:

$$p_{n+1} = (\rho + 1)p_n - \rho p_{n-1}, \quad n \geq 1$$

- Define a linear operator D on $\{a_0, a_1, a_2, \dots\}$ by

$$Da_n = a_{n+1}, \quad \forall n \quad \Longrightarrow D^m a_n = a_{n+m}, \quad \forall n, m$$

- A general linear difference equation with constant coefficients

$$C_n a_n + C_{n+1} a_{n+1} + \cdots + C_{n+k} a_{n+k} = 0$$

can be written as

$$C_n a_n + C_{n+1} D a_n + \cdots + C_{n+k} D^k a_n = 0$$

How to solve, we will take a simpler second-order case. So, what do we have as a second-order?

Example. *(Solving a second-order linear difference equation)*

- A second-order difference equation of the form $C_2 a_{n+2} + C_1 a_{n+1} + C_0 a_n = 0$ can be written as

$$\left(C_2 D^2 + C_1 D + C_0\right) a_n = 0$$

The corresponding characteristic equation is

$$C_2 r^2 + C_1 r + C_0 = 0$$

The roots of this equation determine the form of the solution for $\{a_n\}$.

- Suppose that r_1 and r_2 are two distinct real roots of the characteristic equation. Then we know r_1^n is a solution to this difference equations, and any constant times of that would also be a solution to that.

$$a_n = d_1 r_1^n \quad \textit{and} \quad a_n = d_2 r_2^n, \quad d_1, d_2 \textit{ are constants}$$

are both solutions of the second order linear difference equation.

- $a_n = d_1 r_1^n + d_2 r_2^n$ is the most general solution.
- *Note the similarity with the theory of solving linear differential equations!*

Now, this idea is what is the difference equations and how one can solve them.

- For our $M/M/1$ model, the main equation in the balance equations can be written as

$$\mu p_{n+2} - (\lambda + \mu) p_{n+1} + \lambda p_n = 0, \quad n \geq 0.$$

Then $\{p_n\}$ are the solution to $\left[\mu D^2 - (\lambda + \mu) D + \lambda\right] p_n = 0$ subject to the boundary conditions $p_1 = \dfrac{\lambda}{\mu} p_0 = \rho p_0$ and $\displaystyle\sum_{n=0}^{\infty} p_n = 1$.

The characteristic equation $\mu r^2 - (\lambda + \mu) r + \lambda = 0$ yields two roots given by 1 and λ/μ and hence the general solution is given by

$$p_n = d_1 (1)^n + d_2 \left(\frac{\lambda}{\mu}\right)^n = d_1 + d_2 \rho^n$$

Now, $\displaystyle\sum_{n=0}^{\infty} p_n = 1 \implies d_1 = 0$ (otherwise $\sum_{n=0}^{\infty} p_n$ would be infinite).
And $p_1 = \rho p_0 \implies d_2 = p_0$ (since $p_1 = d_2 \rho$ from the above).
Thus, $p_n = p_0 \rho^n$ and p_0 is found to be $p_0 = 1 - \rho$ for $\rho < 1$, giving us the required result.

So, these are three different methods of solving this particular $M/M/1$ system.

- For $M/M/1$ queue, the steady state system size probabilities are given by a geometric distribution.
- p_n depends on λ and μ only through $\rho = \lambda/\mu$.
- The steady state solution for the $M/M/1$ system exists under the condition that $\rho < 1$, or equivalently $\lambda < \mu$, or equivalently $\frac{1}{\mu} < \frac{1}{\lambda}$ (intuitive!).
  Q: Why no equilibrium solution when $\lambda = \mu$?

  We can see that when $\lambda = \mu$ also, exactly, it would match, but exactly it would match provided this is there is no randomness here; because of this randomness, there may be patches at which the system will be empty. And that will not be able to serve because the mean service rate is exactly equal to the mean arrival rate. So, it will not be able to serve completely, and then the queue will start building up so that the stability is lost.
- $\rho < 1$ is the condition for ergodicity of the process and the stability condition for the $M/M/1$ system.
- The steady state probability distribution for the system size $p_n = P\{N = n\} = (1-\rho)\rho^n, n = 0, 1, 2, \ldots, \rho = \frac{\lambda}{\mu} < 1$ allows us to calculate various equilibrium performance measures of the system.

So, we will stop here at this point after having obtained the equilibrium system size distributions for an $M/M/1$ queue. So, we will continue with performance measures in the next lecture.
Thank you. Bye.